Transliteration Extraction from Classical Chinese Buddhist Literature Using Conditional Random Fields

Yu-Chun Wang
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
Telecommunication Laboratories, Chunghwa Telecom, Taiwan
d97023@csie.ntu.edu.tw

Richard Tzong-Han Tsai*
Department of Computer Science and Information Engineering, National Central University, Zhongli City, Taiwan
thtsai@csie.ncu.edu.tw

Abstract

Extracting plausible transliterations from historical literature is a key issues in historical linguistics and other resaech fields. In Chinese historical literature, the characters used to transliterate the same loanword may vary because of different translation eras or different Chinese language preferences among translators. To assist historical linguiatics and digial humanity researchers, this paper propose a transliteration extraction method based on the conditional random field method with the features based on the characteristics of the Chinese characters used in transliterations which are suitable to identify transliteration characters. To evaluate our method, we compiled an evaluation set from the two Buddhist texts, the Samyuktagama and the Lotus Sutra. We also construct a baseline approach with suffix array based extraction method and phonetic similarity measurement. Our method outperforms the baseline approach a lot and the recall of our method achieves 0.9561 and the precision is 0.9444. The results show our method is very effective to extract transliterations in classical Chinese texts.

1 Introduction

Cognates and loanwords play important roles in the research of language origins and cultural interchange. Therefore, extracting plausible cognates or loanwords from historical literature is a key issues in historical linguistics. The adoption of loanwords from other languages is usually through transliteration. In Chinese historical literature, the characters used to transliterate the same loanword may vary because of different translation eras or different Chinese language/dialect preferences among translators. For example, in classical Chinese Buddhist scriptures, the translation process of Buddhist scriptures from Sanskrit to classical Chinese occurred mainly from the 1st century to 10th century. In these works, the same Sanskrit words may transliterate into different Chinese loanword forms. For instance, the surname of the Buddha, Gautama, is transliterated into several different forms such as "瞿曇" (*qü-tan*) or "喬答摩" (*qiao-da-mo*), and the name "Culapanthaka" has several different Chinese transliterations such as "朱利槃特" (*zhu-li-pan-te*) and "周利槃陀伽" (*zhou-li-pan-tuo-qie*). In order to assist researchers in historical linguistics and other digital humanity research fields, an approach to extract transliterations in classical Chinese texts is necessary.

Many transliteration extraction methods require a bilingual parallel corpus or text documents containing two languages. For example, (Sherif and Kondrak, 2007) proposed a method for learning the string distance measurement function from a sentence-aligned English-Arabic parallel corpus to extract transliteration pairs. (Kuo et al., 2007) proposed a transliteration pair extraction method using a phonetic similarity model. Their approach is based on the general rule that when a new English term is transliterated into Chinese (in modern Chinese texts, e.g. newswire), the English source term usually appears alongside the transliteration. To exploit this pattern, they identify all the English terms in a Chinese text and measure the phonetic similarity between those English terms and their surrounding Chinese terms, treating the pairs with the highest similarity as the true transliteration pairs. Despite its high accuracy, this approach cannot be applied to transliteration extraction in classical Chinese literature since the prerequisite (of the source terms alongside the transliteration) does not apply.

27th Pacific Asia Conference on Language, Information, and Computation pages 260－266

Some researchers have tried to extract transliterations from a single language corpus. (Oh and Choi, 2003) proposed a Korean transliteration identification method using a Hidden Markov Model (HMM) (Rabiner, 1989). They transformed the transliteration identification problem into a sequential tagging problem in which each Korean syllable block in a Korean sentence is tagged as either belonging to a transliteration or not. They compiled a human-tagged Korean corpus to train a hidden Markov model with predefined phonetic features to extract transliteration terms from sentences by sequential tagging. (Goldberg and Elhadad, 2008) proposed an unsupervised Hebrew transliteration extraction method. They adopted an English-Hebrew phoneme mapping table to convert the English terms in a named entity lexicon into all the possible Hebrew transliteration forms. The Hebrew transliterations are then used to train a Hebrew transliteration identification model. However, Korean and Hebrew are alphabetical writing system, while Chinese is ideographic. These identification methods heavily depend on the phonetic characteristics of the writing system. Since Chinese characters do not necessarily reflect actual pronunciations, these methods are difficult to apply to the transliteration extraction problem in classical Chinese.

This paper proposes an approach to extract transliterations automatically in classical Chinese texts, especially Buddhist scriptures, with supervised learning models based on the probability of the characters used in transliterations and the anguage model features of Chinese characters.

2 Method

To extract the transliterations from the classical Chinese Buddhist scriptures, we adopt a supervised learning method, the conditional random fields (CRF) model. The features we use in the CRF model are described in the following subsections.

2.1 Probability of each Chinese character in transliterations

According to our observation, in the classical Chinese Buddhist texts, the Chinese characters chosen be used in transliteration show some characteristics. Translators tended to choose the characters that do not affect the comprehension of the sentences. The amount of the Chinese characters is huge, but the possible syllables are limited in Chinese. Therefore, one Chinese character may share the same pronunciation with several other characters. Hence, the translators may try to choose the rarely used characters for transliteration.

From our observation, the probability of each Chinese character used to be transliterated is an important feature to identify transliteration from the classical Buddhist texts. In order to measure the probability of every character used in transliterations, we collect the frequency of all the Chinese characters in the Chinese Buddhist Canon. Furthermore, we apply the suffix array method (Manzini and Ferragina, 2004) to extract the terms with their counts from all the texts of the Chinese Buddhist Canon. Next, the extracted terms are filtered out by the a list of selected transliteration terms from the Buddhist Translation Lexicon and Ding Fubao's Dictionary of Buddhist Studies. The extracted terms in the list are retained and the frequency of each Chinese character can be calculated. Thus, the probability of a given Chinese character c in transliteration can be defined as:

$$Prob(c) = \log \frac{freq_{trans}(c)}{freq_{all}(c)}$$

where $freq_{trans}(c)$ is c's frequency used in transliterations, and $freq_{all}(c)$ is c's frequency appearing in the whole Chinese Buddhist Canon. The logarithm in the formula is designed for CRF discrete feature values.

2.2 Language model of the transliteration

Transliterations may appear many times in one Buddhist sutra. The preceding character and the following character of the transliteration may be different. For example, for the phrase "於憍薩羅國" (*yu-jiao-sa-luo-guo*, in Kosala state), if we want to identify the actual transliteration, "憍薩羅" (*jiao-sa-luo*, Kosala), from the extra characters "於" (*yu*, in) and "國" (*guo*, state), we must first use an effective feature to identify the boundaries of the transliteration.

In order to identify the boundaries of transliterations, we propose a language-model-based feature. A language model assigns a probability to a sequence of m words $P(w_1, w_2, \ldots, w_m)$ by means of a probability distribution. The probability of a sequence of m words can be transformed

into a conditional probability:

$$\begin{aligned} P(w_1, w_2, \cdots, w_m) &= P(w_1)P(w_2|w_1) \\ &\quad P(w_3|w_1, w_2) \cdots \\ &\quad P(w_m|w_1, w_2, \cdots, \\ &\quad w_{m-1}) \\ &= \prod_{i=1}^{m} P(w_i|w_1, w_2, \cdots, \\ &\quad w_{i-1}) \end{aligned}$$

In practice, we can assume the probability of a word only depends on its previous word (bi-gram assumption). Therefore, the probability of a sequence can be approximated as:

$$\begin{aligned} P(w_1, w_2, \cdots, w_m) &= \prod_{i=1}^{m} P(w_i|w_1, w_2, \\ &\quad \cdots, w_{i-1}) \\ &\approx \prod_{i=1}^{m} P(w_i|w_{i-1}) \end{aligned}$$

We collect person and location names from the Buddhist Authority Database[1] and the known Buddhist transliteration terms from The Buddhist Translation Lexicon (翻譯名義集)[2] to create a dataset with 4,301 transliterations for our bi-gram language model.

After building the bi-gram language model, we apply it as a feature for the supervised model. Following the previous example, "於憍薩羅國" (*yu-jiao-sa-luo-guo*, in Kosala state), for each character in the sentence, we first compute the probability of the current character and its previous character. For the first character "於" , since there is no previous word, the probability is $P(於)$. For the second character "憍", the probability of the two characters is $P(於憍) = P(於)P(憍|於)$. We then compute the probability of the second and third characters: $P(憍薩) = P(憍)P(薩|憍)$, and so on. If the probability changes sharply from that of the previous bi-gram, the previous bi-gram may be the boundary of the transliteration. Because the character "於" rarely appears in transliterations, $P(於憍)$ is much lower than $P(憍薩)$. We may conclude that the left boundary is between the first two characters "於憍".

2.3 Functional Words

We take the classical Chinese functional words into consideration. These characters have special grammatical functions in classical Chinese; thus, they are seldom used to transliterate foreign names. This is a binary feature which records the character is a functional word or not. The functional words are listed as follows: 之 (*zhi*), 乎 (*hu*), 且 (*qie*), 矣 (*yi*), 邪 (*ye*), 於 (*yu*), 哉 (*zai*), 相 (*xiang*), 遂 (*sui*), 嗟 (*jie*), 與 (*yu*), and 噫 (*yi*).

2.4 Appellation and Quantifier Words

After observing the transliterations appearing in classical Chinese literature, we note that there are some specific patterns of the characters follows the transliteration terms. Most of the characters following the transliteration are appellation or quantifier words, such as 山 (*san*, mountain), 海 (*hai*, sea), 國 (*guo*, state), 洲 (*zhou*, continent). For example, there are some cases like 耆闍崛山 (*qi-du-jui-san*, Vulture mountain), 拘薩羅國 (*jü-sa-luo-guo*, Kosala state), and 贍部洲 (*zhan-bu-zhou*, Jambu continent). Therefore, we collect the Chinese characters that are usually used as appellation or quantifiers following transliterations and then design this feature. This is also a binary feature that records the character is used as an appellation or quantifier word or not.

2.5 CRF Model Training

We adopt the supervised learning models, conditional random field (CRF) (Lafferty et al., 2011), to extract the transliterations in classical Buddhist texts. For CRF model, we formulate the transliteration extraction problem as a sequential tagging problem.

2.5.1 Conditional Random Fields

Conditional random fields (CRFs) are undirected graphical models trained to maximize a conditional probability (Lafferty et al., 2011). A linear-chain CRF with parameters $\Lambda = \lambda_1, \lambda_2, \ldots$ defines a conditional probability for a state sequence $\mathbf{y} = \mathbf{y}_1 \ldots \mathbf{y}_T$, given that an input sequence $\mathbf{x} = \mathbf{x}_1 \ldots \mathbf{x}_T$ is

$$P_\Lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\mathbf{x}} \exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{x}, t)\right)$$

where $Z_\mathbf{x}$ is the normalization factor that makes the probability of all state sequences sum to one; $f_k(\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{x}, t)$ is often a binary-valued feature function and λ_k is its weight. The feature functions can measure any aspect of a state transition, $\mathbf{y}_{t-1} \rightarrow \mathbf{y}_t$, and the entire observation sequence,

[1] http://authority.ddbc.edu.tw/

[2] http://www.cbeta.org/result/T54/T54n2131.htm

$\mathbf{x}$, centered at the current time step, t. For example, one feature function might have the value 1 when $\mathbf{y}_{t-1}$ is the state *B*, $\mathbf{y}_t$ is the state *I*, and $\mathbf{x}_t$ is the character "國" (*guo*). Large positive values for λ_k indicate a preference for such an event; large negative values make the event unlikely.

The most probable label sequence for $\mathbf{x}$,

$$\mathbf{y}^* = \arg\max_y P_\Lambda(\mathbf{y}|\mathbf{x})$$

can be efficiently determined using the Viterbi algorithm.

2.5.2 Sequential Tagging and Feature Template

The classical Buddhist texts are separated into sentences by the Chinese punctuation. Then, each character in the sentences is taken as a data row for CRF model. We adopt the tagging approach motivated by the Chinese segmentation (Tsai et al., 2006) which treat Chinese segmentation as a tagging problem. The characters in a sentence are tagged in **B** class if it is the first character of a transliteration word or in **I** class if it is in a transliteration word but not the first character. The characters that do not belong to a transliteration words are tagged in **O** class. We adopt the CRF++ open-source toolkit[3]. We train our CRF models with the unigram and bigram features over the input Chinese character sequences. The features are shown as follows.

- Unigram: s_{-2}, s_{-1}, s_0, s_1, s_2
- Bigram: $s_{-1}s_0$, s_0s_1

where current substring is s_0 and s_i is other characters relative to the position of the current character.

3 Evaluation

3.1 Data set

We choose one Buddhist scripture as our data set for evaluation from the Chinese Buddhist Canon maintained by Chinese Buddhist Electronic Text Association (CBETA). The scripture we choose to compile the training and test sets is the Samyuktagama (雜阿含經). The Samyuktagama is one of the most important scriptures in Early Buddhism and contains a lot of transliterations because it detailedly records the speech and the lives of the Buddha and many of his diciples.

[3] http://crfpp.googlecode.com

The Samyuktagama is an early Buddhist scripture collected shortly after the Buddha's death. The term agama in Buddhism refers to a collection of discourses, and the name Samyuktagama means "connected discourses." It is among the most important sutras in Early Buddhism. The authorship of the Samyuktagama is traditionally regarded as the most early sutra collected by the Mahakssyapa, the Buddha's disciple, and five hundred Arhats three months after the Buddha's death. An Indian monk, Gunabhadra, translated this sutra into classical Chinese in Liu Song dynasty around 443 C.E. The classical Chinese Samyuktagama has 50 volumes containing about 660,000 characters. Because the amount of Samyuktagama is too tremendous, we take the first 20 volumes as the training set, and the last 10 volumes as the test set.

In addition, we want to evaluate if the supervised learning model trained by one Buddhist scripture can be applied to another Buddhist scrupture translated in different era. Therefore, we choose another scripture, the Lotus Sutra (妙法蓮華經), to create another test set. The Lotus sutra is a famous Mahayana Buddhist scripture probably written down between 100 BC and 100 C.E. The earliest known Sanskrit title for the sutra is the Saddharma Pundarika Sutra, which translates to "the Good Dharma Lotus Flower Sutra." In English, the shortened form Lotus Sutra is common. The Lotus Sutra has also been highly regarded in a number of Asian countries where Mahayana Buddhism has been traditionally practiced, such as China, Japan, and Korea. The Lotus Sutra has several classical Chinese translation versions. The most widely used version is translated by Kumarajiva ("鳩摩羅什" in Chinese) in 406 C.E. It has eight volumes and 28 chapters containing more then 25,000 characters. We select the first 5 chapters as a different test set to evaluate our method.

3.2 Baseline Method

There are a few reseaches focusing on transliteration extraction from classical Chinese literature. However, in order to compare and show the benefits of our method, we construct a baseline system with widely used information extraction methods. Because many previous researches on transliteration extraction are based on phonetic similarity or phoneme mapping approaches, we also use these methods to construct the baseline system. First,

Table 1: Evaluation Results of Tranliteration Extraction

		Precision	Recall	F_1-score
Our Approach	The Samyuktagama test set	0.8810	0.9561	0.9170
	The Lotus Sutra test set	0.9444	0.9474	0.9459
Baseline	The Samyuktagama test set	0.0399	0.7771	0.0759
	The Lotus Sutra test set	0.0146	0.5789	0.2848

the baseline system use the suffix array method to extract all the possible terms for the classical Chinese Buddhsit scriptures. Then, the extracted terms are converted into Pinyin sequences by a modern Chinese pronunication dictionary. We also adopt the collected transliteation list used in section 2.1 and also convert the transliterations into Pyinyin sequences. Next, for each extracted terms, the baseline system measures the Levenshtein distance between the Pinyin sequences of the extracted terms and all the transliterations as the phonetic similairy. If the extracted term has a Levenshtein distance less than threshold (distance $\leq$ 3 in our baseline) from one of the transliterations we collect, the extracted term will be regarded as a transliteration; otherwise, the term will be dropped.

3.3 Evaluation Metrics

We use two evaluation metrics, recall and precision, to estimate the performance of our system. Recall and precision are widely used measurements in many research fields, sucn as information retrieval and information extraction. (Manning et al., 2008) In the digital humanities research field, a key issue is the coverage of the extraction method. To maximize usefulness to researchers, a method should be able to extract as many potential transliterations from literature as possible. Therefore, in our evaluation, we use recall, defined as follows:

$$Recall = \frac{|\text{Correctly extracted transliterations}|}{|\text{Transliterations in the data set}|}$$

In addition, the correctness of the extracted transliterations are also important. To avoid wasting time on the useless information, a method should be able to extract correct transliterations from literature as possible. Thus, we also use precision, defined as follows:

$$Precision = \frac{|\text{Correctly extracted transliterations}|}{|\text{All extracted transliterations}|}$$

With precision and recall, the F-score measurement is also adopted as a weighted average of the precision and recall. The F_1-score is defineds as follows:

$$F_1\text{-score} = \frac{2 \times recision \times recall}{precision + recall}$$

3.4 Evaluation Results

Table 1 shows the results of our method and the baseline system on different test sets. The gold standards of these two test sets are compiled by human experts who examine all the sentences in the test sets and regconize each transliterations for evaluation. The results show that our method can extract 95.61% transliterations on the Sumyuktagama and 94.74% on the Lotus Sutra. On the precision measurement, our method also achieves pretty good results, which show that most of the terms our method extract are actual transliterations. Our method outperforms the baseline system and the precision of the baseline system is very poor. The baseline system cannot extract most transliterations due to the limit of the suffix array method since the suffix array method only extracts the terms that appear twice or more in the context. Besides, the phonetic similarity is not effective to filter the transliteartions; the problem causes the low precision. These results demonstrate that our method can save a lot of labor-intensive work to examine the transliteration for the historical and humanity researchers.

4 Discussion

4.1 Effectiveness of transliteration extraction

Our method can extract many transliterations from the Samyuktagama such as "迦毘羅衛" (*jia-pi-luo-wei*, *Kapilavastu*, the name of an ancient kingdom where the Buddha was born and grew up), "尼拘律" (*ni-jü-lü*, *Nyagro*, the forest name in Kapilavastu kingdom), and "摩伽陀" (*muo-qie-tuo*, *Magadha*, the name of an ancient Indian kingdom). These transliteration do not appear in the training set, but our method can still identify them. In addition, our method also finds out many transliterations in the Lotus Sutra which

are unseen in the Samyuktagama, such as "娑伽羅" (*suo-qie-luo*, *Sagara*, the name of the king of the sea world in ancient Indian mythology), "鳩槃荼/鳩槃茶" (*jiu-pan-cha*/*jiu-pan-tu*, *Kumbhanda*, one of a group of dwarfish, misshapen spirits among the lesser deities of Buddhist mythology), and "阿鞞跋致" (*a-pi-ba-zhi*, *Avaivart*, "not turn back" in Sanskrit). Since the characteristics of the Lotus Sutra are different from the Samyuktagama in many aspects, it shows that the supervised learning model trained by one Buddhist scripture may apply to other Buddhist scriptures translated in different eras and translators.

We also discovered that transliterations may vary even in the same scripture. In the Samyuktagama, the Sanskrit term "Chandala" (someone who deals with disposal of corpses, and is a Hindu lower caste, formerly considered untouchables) has two different transliterations: "旃陀羅" (*zhan-tuo-luo*) and "栴陀羅" (*zhan-tuo-luo*). The Sanskrit term "*Magadha*" (the name of an ancient Indian kingdom) has three different transliterations: "摩竭陀" (*muo-jie-tuo*), "摩竭提" (*muo-jie-ti*), and "摩伽陀" (*muo-qie-tuo*). The variations of the transliterations of the same word give the clues of translators and translation progress. These variations may help the study of historical Chinese phonology and philology.

4.2 Error cases

Although our method can extract and identify most transliteration pairs, some transliteration pairs cannot be identified. The error cases can be divided into several categories. The first one is that a few terms cannot be extracted, such as "闍維" (*she-wei*, *Jhapita*, cremation, a monk's funeral pyre). This transliteration is less used and only appears three times in the final part of the Samyuktagama. The widely used transliteration of the term "*Jhapita*" is "荼毘" (*tu-pi*). It may cause the difficulty for the supervised learning model to identify these terms.

The other case is incorrect boundary of the transliterations. Sometimes our method may extract shorter terms, such as "韋提" (*wei-ti*, correct transliteration is "韋提希", *wei-ti-xi*, *Vaidehi*, a female person name), "波羅" (*po-luo*, correct transliteration is "波羅柰", *po-luo-nai*, *Varanasi*, a location name in northen India), "瞿利摩羅" (*qü-li-muo-luo*, correct transliteration is "央瞿利摩羅", *yang-qü-li-muo-luo*, *Angulimala*, one of the Buddha's disciples). This problem is due to the probability generated by the language model. For example, the probability of the first two charactgers of the transliteration "央瞿利摩羅", $P(\text{央瞿})$, is very low. It causes the CRF model predicts the first character "央" (*yang*) does not belong to the transliteration. If more transliterations can be collected to build a better language model, this problem can be overcome.

In some cases, our method extracts much longer terms, like "阿那律陀夜" (*a-na-lü-tuo-ye*, correct transliteration is "阿那律陀", *a-na-lü-tuo*, *Aniruddha*, one of the Buddha's closest disciples), and "兒富那婆藪" (*er-fu-na-po-sou*, correct transliteration is "富那婆藪", *fu-na-po-sou*, *Punabbasu*, a kind of ghost in Buddhist mythology). In these cases, the previous or following characters are often used in transliterations. Therefore, it is very difficult to distinguish the boundary of the actual transliteration. In addition, there are some cases that a transliteration followed by another transliteration immediately. For example, our method extracts out the term "闡陀舍利" (*chan-tuo-she-li*), which comprises two transliteration terms such as "闡陀" (*chan-tuo*, *Chanda*, one of the Buddhist's disciples) and "舍利" (*she-li*, *Sarira*, Buddhist relics). It is also difficult to separate them without any additional semantic clues. Although our method sometimes might extract incomplete transliterations with incorrect boundary, checking the boundary of a transliteration is not difficult to a human expert. Therefore, the extracted incorrect transliteartions also have the benefits to help humanity researchers quickly find and check plausible transliterations.

5 Conclusion

The transliteration extrction of foreign loanwords is an important task in research fields such as historical linguistics and digital humanities. We propose an approach which can extract transliteration automatically from classical Chinese Buddhist scriptures. Our approach comprises the conditional random fields method with designed features which are suitable to identify transliteration characters. The first feature is the probability of each Chinese character used in transliterations. The second feature is probability of the sequential bigram characters measured by the language model method. In addiition, the functional words, appellation and quantifier words also be regarded

as binary features. Next, the transliteration extration problem is formulated as a sequential tagging problem and the CRF method is used to train a model to extract the transliterations from the input classical Chinese sentences. To evaluate our method, we constructed an evaluation set from the two Buddhist texts, the Samyuktagama and the Lotus Sutra, which were translated into Chinese in different eras. We also construct a baseline system with proach with suffix array based extraction method and phonetic similarity measurementfor comparison. The recall of our method achieves 0.9561 and the precision is 0.9444. The results show our method outperforms the baseline system a lot and is effective to extract transliterations from classical Chinese texts. Our method can find the transliterations among the immense classical literatures to help many research fields such as historical linguistics and philology.

References

Y. Goldberg and M. Elhadad. 2008. Identification of transliterated foreign words in hebrew script. *Computational Linguistics and Intelligent Text Processing*.

J-S. Kuo, H. Li, and Y-K. Yang. 2007. A phonetic similarity model for automatic extraction of transliteration pairs. *ACM Trans. Asian Language Information Processing*, 6(2).

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2011. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. of ICML*, pages 282–289.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

G. Manzini and P. Ferragina. 2004. Engineering a lightweight suffix array construction algorithm. *Algorithmica*, 40(1):33–50.

J. Oh and K. Choi. 2003. A statistical model for automatic extraction of korean transliterated foreign words. *International Journal of Computer Processing of Oriental Languages*, 16(1):41–62.

L. Rabiner. 1989. tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77.

T. Sherif and G. Kondrak. 2007. Bootstrapping a stochastic transducer for arabic-english transliteration extraction. *Proceedings of Annual Meeting- Association for Computational Linguistics*.

Richard Tzong-Han Tsai, Hsieh-Chuan Hung, Cheng-Lung Sung, Hong-Jie Dai, and Wen-Lian Hsu. 2006. On closed task of chinese word segmentation: An improved crf model coupled with character clustering and automatically generated template matching. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 134–137.